Report Analisi dati Covid-19

Gruppo horror

Introduzione

Il lavoro è stato svolto con l'obiettivo di analizzare l'impatto del COVID-19 sul territorio italiano.

Le analisi sono state condotte a livello nazionale, regionale e comunale, analizzando i tassi di mortalità e di contagio e il loro andamento nel tempo.

Una sezione di approfondimento ha riguardato analisi di correlazione tra i tassi di mortalità e l'inquinamento dell'aria e dell'acqua.

Nella parte finale, sono stati inseriti spunti su possibili analisi future, con l'obiettivo di analizzare l'impatto del virus sull'andamento scolastico, sui disturbi mentali e in relazione con i paesi europei.

1. Presentazione e pulizia dei dataset

La pulizia dei 4 dataset Regioni, Province, Comuni, Ripartizione Geografica (di cui i primi 3 files csv, Comuni trasformato da file xls a csv, l'ultimo file di testo) è stata effettuata su Excel, tramite lo strumento "Power Query", dove sono stati rimossi i valori nulli, le celle vuote, eventuali duplicati e gli errori. Sono state poi rimosse alcune colonne non utili ai fine dell'analisi (ad es. i valori di longitudine e latitudine). Il lavoro di pulizia ha riguardato anche i nomi di alcune regioni, allo scopo di uniformare tutti i dati: c'erano infatti regioni con denominazione diversa tra i vari file (es, Valle d'Aosta, Trentino-Alto Adige).

Volendo approfondire ciascun dataset, si osserva che:

1. il dataset "Comuni" (7909 righe, 4 colonne) contiene i dati relativi ai singoli comuni e i relativi abitanti. È stato utilizzato principalmente per ricavare la popolazione totale;
2. il dataset "Regioni" (6028 righe,12 colonne) è stato utilizzato per la maggior parte delle analisi, perché contenente tutti i dati relativi alle ospedalizzazioni, i contagi nuovi e correnti, i decessi, i guariti e il numero totale di positivi. I dati sono stati rilevati nell'arco temporale che va dal 24/02/2020 al 06/12/2020. Di questi dati, il numero di decessi giornalieri è stato raccolto in maniera cumulativa. Per fare un'analisi sul numero assoluto di decessi giornalieri, è stata aggiunta una nuova colonna al dataset, calcolando la differenza con il giorno precedente.
   Inoltre, alcuni dati contengono valori aggregati, nello specifico:
   - "TotalHospitalizedPatients" comprende il numero dei pazienti ospedalizzati e dei pazienti in cura intensiva;
   - "CurrentPositiveCases" include il numero dei positivi in isolamento e il totale dei pazienti in ospedale;
   - "TotalPositiveCases" aggrega la somma del numero di positivi, guariti e decessi.
3. Nel dataset "Province" (40202 righe, 7 colonne), sono presenti i dati sulle province e i casi di contagio giornalieri cumulativi;
4. Infine, il dataset delle "Ripartizioni Geografiche" (21 righe, 3 colonne) contiene il nome delle regioni con le corrispettive aree di appartenenza (NordOvest, NordEst, Centro, Sud, Isole).

Si passa ora ad approfondire le analisi effettuate su Python, dove sono stati importati i dataset ed aperti tramite la libreria Pandas. I grafici sono stati realizzati utilizzando le librerie Matplot, Seaborn e Geoplot; inoltre sono stati utilizzate le librerie Pandas e Numpy.

Analisi esplorative

Si calcolano le statistiche descrittive del dataset “Regioni”.

	Ospedalizzati	Cura Intensiva	Totale Ospedalizzati
count	6027	6027	6027
mean	50,9093579	55,892318	564,985897
std	1272,731157	134,957979	1403,362341
min	0	0	0
25%	16	10	17
50%	93	9	106
75%	384,5	46	434
max	12˙077	1˙381	13˙328
total	3˙068˙307	336˙863	3˙405˙170

Nell’arco tempo considerato ci sono stati in media 565 pazienti ricoverati in ospedale, con un picco massimo di 13˙328 pazienti (registrato il giorno 30 aprile), di cui 1˙381 in cura intensiva e 12˙077 solo ricoverati. Anche se la media dei pazienti ricoverati in cura intensiva e semplice ricovero è poco dissimile, osservando i valori mediani si nota come la distribuzione è sicuramente più ampia per i ricoverati semplici.

	Quarantena	Positivi Correnti	Nuovi Positivi
count	6027	6027	6027
mean	5722,911399	6287,897296	286,885847
std	15892,768789	16804,884786	818,450458
min	0	0	-229
25%	139	168	4
50%	743	905	31
75%	3057	3735	159,50
max	155066	164406	11˙489
total	34˙491˙987	37˙897˙157	1˙729˙061

	Guariti	Decessi	Totale Casi Positivi
count	6027	5740	6027
mean	9337,852829	10,466551	17100,12361
std	20551,265039	33,2482201	36934,67627
min	0	-61	0
25%	734	0	1383,5
50%	2840	1	4613
75%	8450	7	15304
max	289706	546	429˙109
total	56˙279˙239	60˙078	103˙062˙445

Il picco dei contagi totali è di 11˙489, registrato in data 7 novembre.

Osservando invece i dati sul totale di casi positivi, più della metà di questo valore comprende i guariti (289˙706), mentre il numero di persone in quarantena e i positivi correnti si differenziano di una migliaia di casi. Dei positivi correnti (37˙897˙157) la maggior parte sono in isolamento (34˙491˙987) e il resto ospedalizzati (3˙405˙170).

In media i decessi sono circa 10 al giorno ma, osservando il valore mediano e confrontandolo col 4° percentile vediamo la distribuzione aumenta notevolmente verso la coda sinistra.

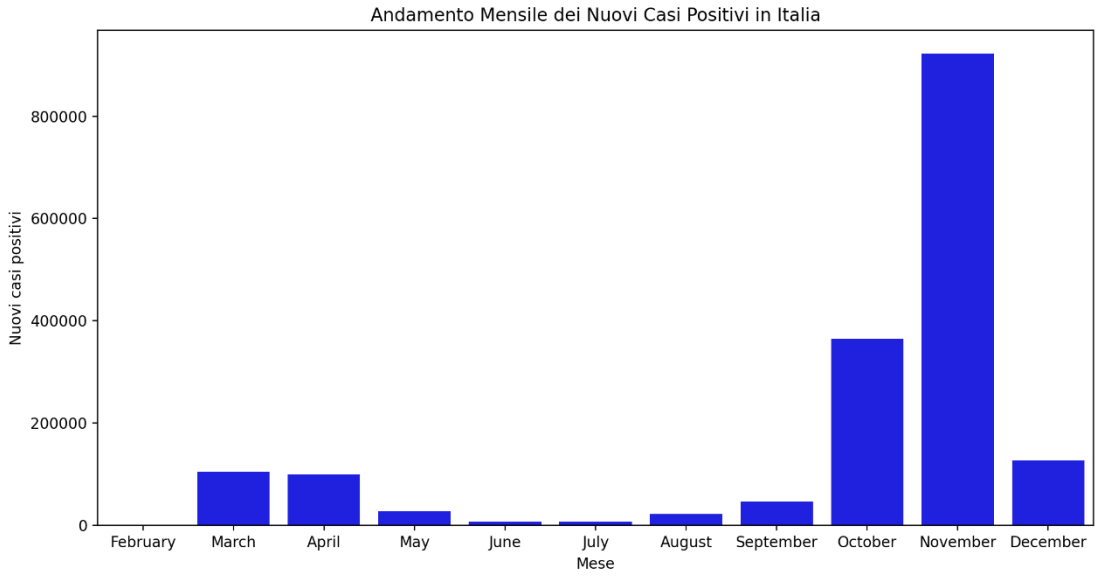
Il picco dei decessi si è registrato in data 4 aprile, con 546 decessi in un giorno.

In conclusione, su una popolazione totale di 59˙422˙954 persone, si sono registrati 1˙729˙061 di contagi in totale, con 60˙078 decessi, con un tasso di mortalità pari allo 0,10%, un tasso di contagio del 2,91% e un tasso di ospedalizzazione del 5,73%.

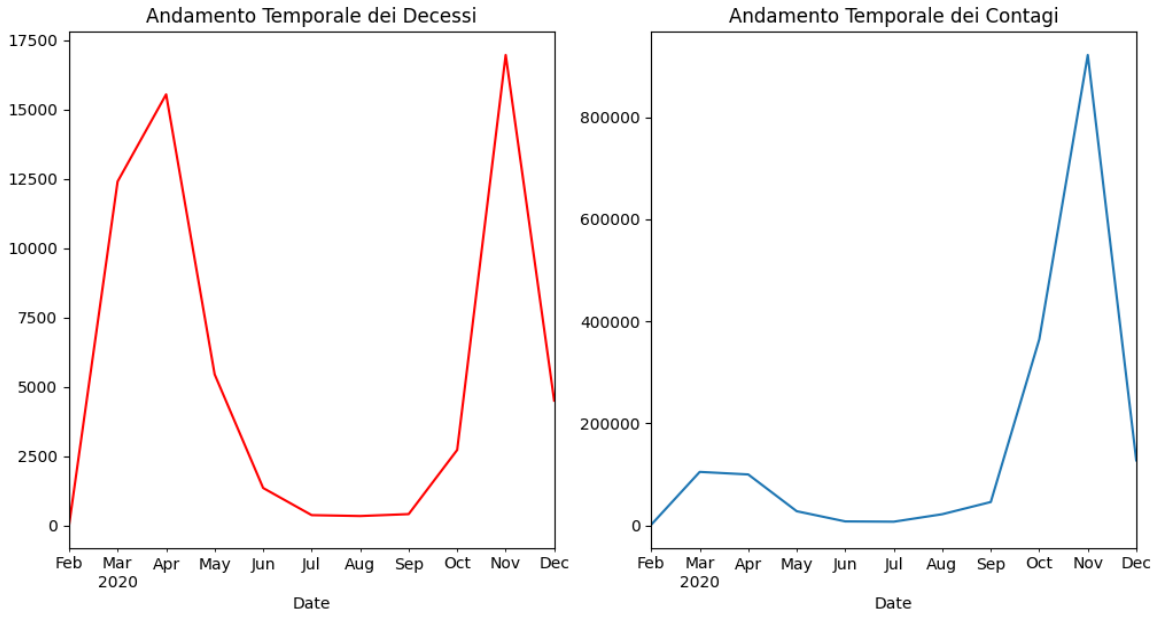
Analisi Regionale

1. Andamento temporale

Dal grafico di seguito si può osservare l'andamento nazionale dei nuovi casi positivi, mese per mese. Si nota come dopo il Dpcm dell'11 marzo 2020, che ha decretato il primo lockdown, il tasso di positività cominci a scendere pian piano, raggiungendo il picco minimo durante i mesi estivi, seguito poi da un'impennata dei casi registrata subito dopo settembre, che coincide con la riapertura delle scuole.

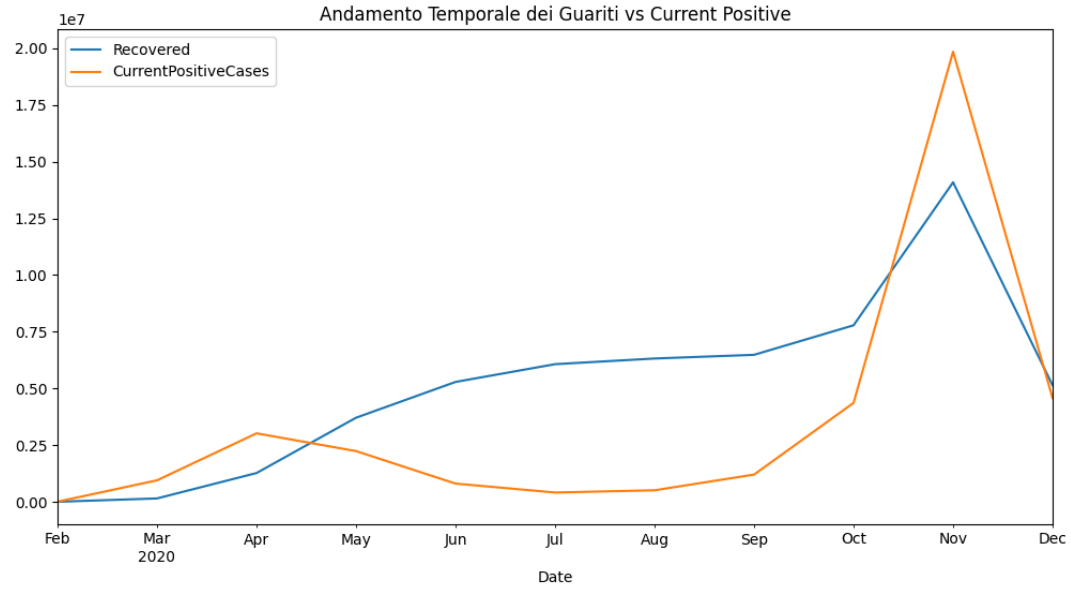


Successivamente è stato analizzato l'andamento temporale, a livello nazionale, dei decessi e messo a confronto con i contagi. Come si può notare il numero di decessi aumenta notevolmente dall'inizio del contagio a febbraio, raggiungendo picchi elevati nei mesi di marzo ed aprile (picco max in data 4 aprile, con 546 decessi in un giorno), dati preoccupanti che hanno influenzato sicuramente le restrizioni imposte dal governo. Il picco dei contagi si è verificato invece a novembre, per la precisione il giorno 7 novembre.



Infine, si è voluto vedere mettere a confronto i dati dei guariti rispetto ai nuovi casi positivi.

Si nota come all'aumentare dei guariti corrisponda una diminuzione del numero dei positivi correnti, una tendenza visibile soprattutto durante i mesi estivi. Un'inversione del pattern si verifica durante novembre, il meso di picco dei contagi, dove il numero dei positivi supera di gran lunga il numero dei guariti/dimessi registrato.

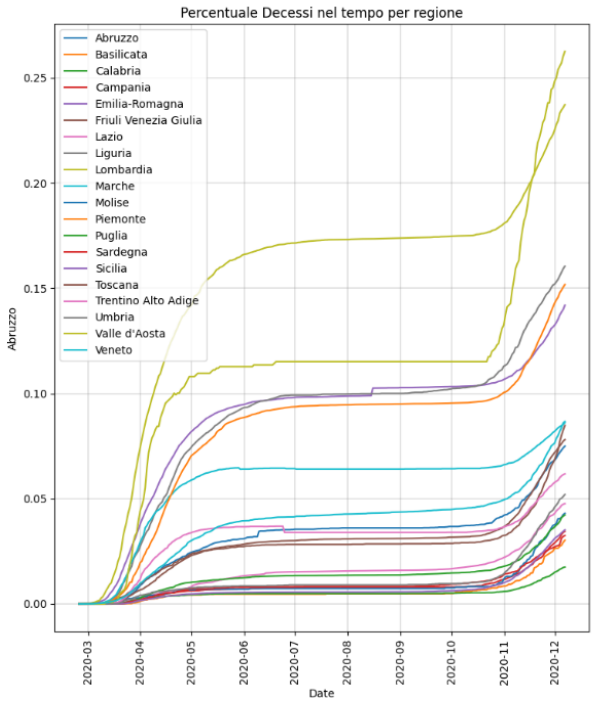


2. L’importanza della normalizzazione

Ai fini di un’analisi più accurata, tutti i dati sono stati normalizzati per la popolazione di riferimento. Prendere come riferimento per le indagini i valori assoluti è una scelta fuorviante, che non riflette il reale andamento del contagio in Italia. Lo si può osservare nelle tabelle e nei grafici di seguito: dando uno sguardo ai dati normalizzati, si può notate che, anche se la Lombardia registra un numero più alti di contagi in valore assoluto, in Valle d’Aosta il contagio ha avuto un impatto maggiore. Il trend si conferma anche per i decessi. Si osservano le sei regioni con il più alto numero di decessi e vediamo come il contagio sia stato di nuovo più incisivo per la Valle d’Aosta.

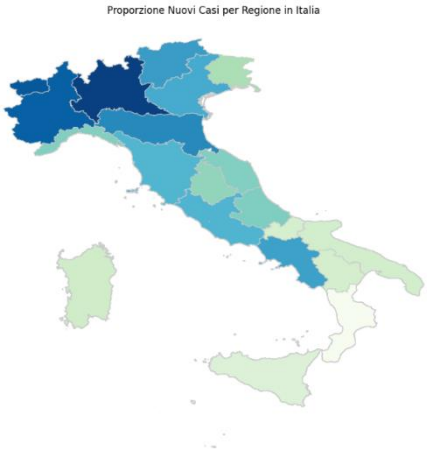
Elenco delle prime sei regioni per numero di contagi			
Dati Non Normalizzati		Dati Normalizzati (%)	
Lombardia	429˙103	Valle d'Aosta	5.30%
Piemonte	177˙788	Lombardia	4.42%
Campania	165˙251	Trentino-Alto Adige	4.10%
Veneto	165˙248	Piemonte	4.07%
Emilia-Romagna	133˙897	Liguria	3.42%
Lazio	130˙254	Veneto	3.40%
Toscana	108˙397	Emilia-Romagna	3.08%

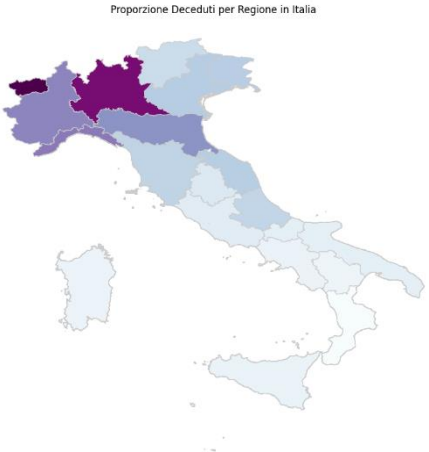
Elenco delle prime sei regioni per numero di decessi			
Dati Non Normalizzati		Dati Normalizzati (%)	
Lombardia	23024	Valle d'Aosta	0.26%
Piemonte	6623	Lombardia	0.23%
Emilia-Romagna	6162	Liguria	0.16%
Veneto	4210	Piemonte	0.15%
Toscana	2867	Emilia-Romagna	0.14%
Lazio	2622	Trentino-Alto Adige	0.12%
Liguria	2521	Veneto	0.09%

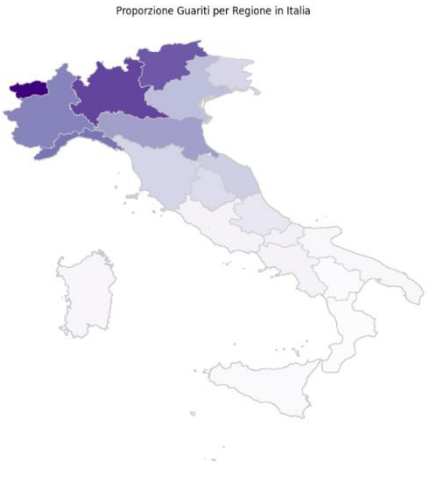


RegionName	Death %
Valle d'Aosta	0.26260587038468214
Lombardia	0.23725929244093583
Liguria	0.1605023002570838
Piemonte	0.1517673575751687
Emilia-Romagna	0.14191175539222065
Veneto	0.08669858382702789
Marche	0.08613616654463502
Friuli Venezia Giulia	0.0848158349115088
Toscana	0.07807304717986646
Abruzzo	0.07496314949258362
Trentino Alto Adige	0.061816017216905536
Umbria	0.052020428195976784
Lazio	0.04764772521182521
Molise	0.043040234648982975
Puglia	0.04227063427926223
Sicilia	0.035159579316333074
Campania	0.03450781281158907
Sardegna	0.032480600334470665
Basilicata	0.030274930973157383
Calabria	0.017508486256093515

I grafici di seguito mostrano una mappa dell'Italia a colori. L'intensità del colore riflette i tassi di mortalità, guarigione e decessi, dove più è intenso il colore, più alti sono i tassi. È evidente come la zona più colpita sia stata il nord-ovest dell'Italia. La Valle d'Aosta ha un elevato tasso di decessi, anche più della Lombardia, anche se la Lombardia è la regione dove si sono registrati il più alto numero di contagi da Coronavirus. Le regioni meno colpite sono le isole e il sud, anche nel sud e nel centro la Campania, il Lazio e l'Emilia-Romagna hanno registrato un tasso elevato di contagio.







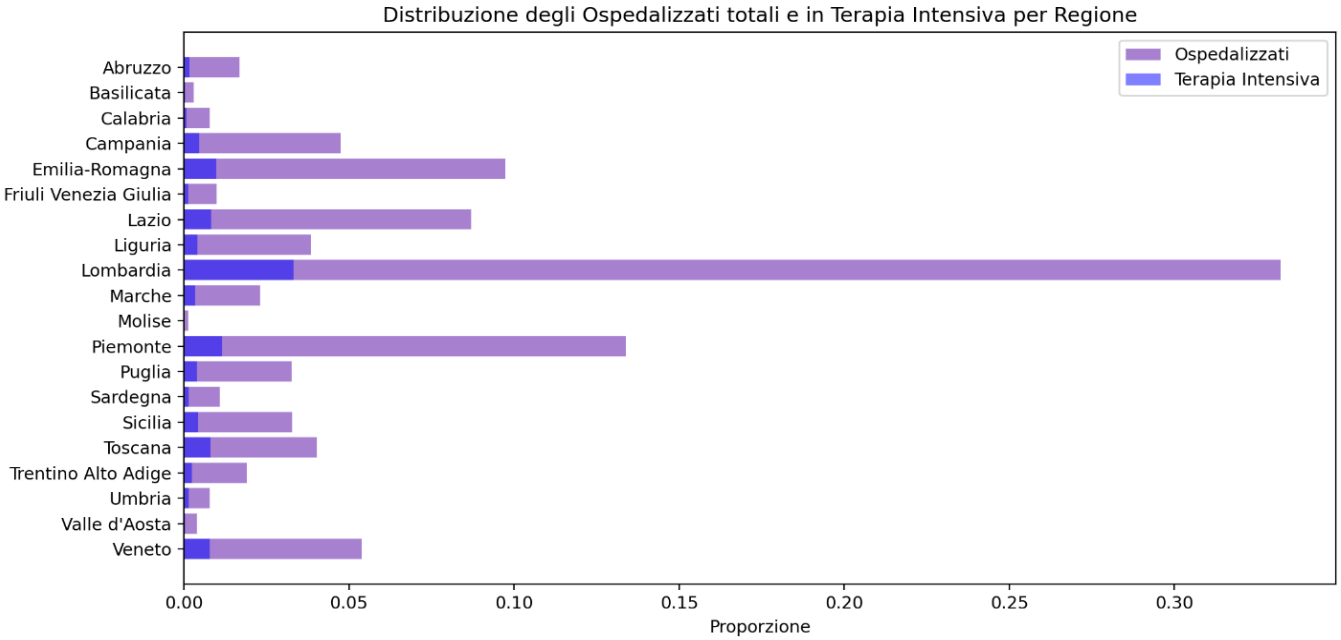
Adesso verranno presentati i dati sulle ospedalizzazioni.

Il grafico rappresenta il tasso di pazienti ricoverati in cura intensiva e semplice ricovero in tutte le regioni d'Italia.

Si può notare una grande differenza tra le due categorie, con una maggiore percentuale di ricoveri semplici a quelli in terapia intensiva. Questo dato si può spiegare in modi diversi:

1. Posti letto in terapia intensiva limitati e non sufficienti ad ospitare i pazienti con una situazione clinica più grave;
2. Un numero minore di casi con quadri clinici gravi rispetto ai casi più lievi ma comunque gravi abbastanza da richiedere il ricovero in ospedale.

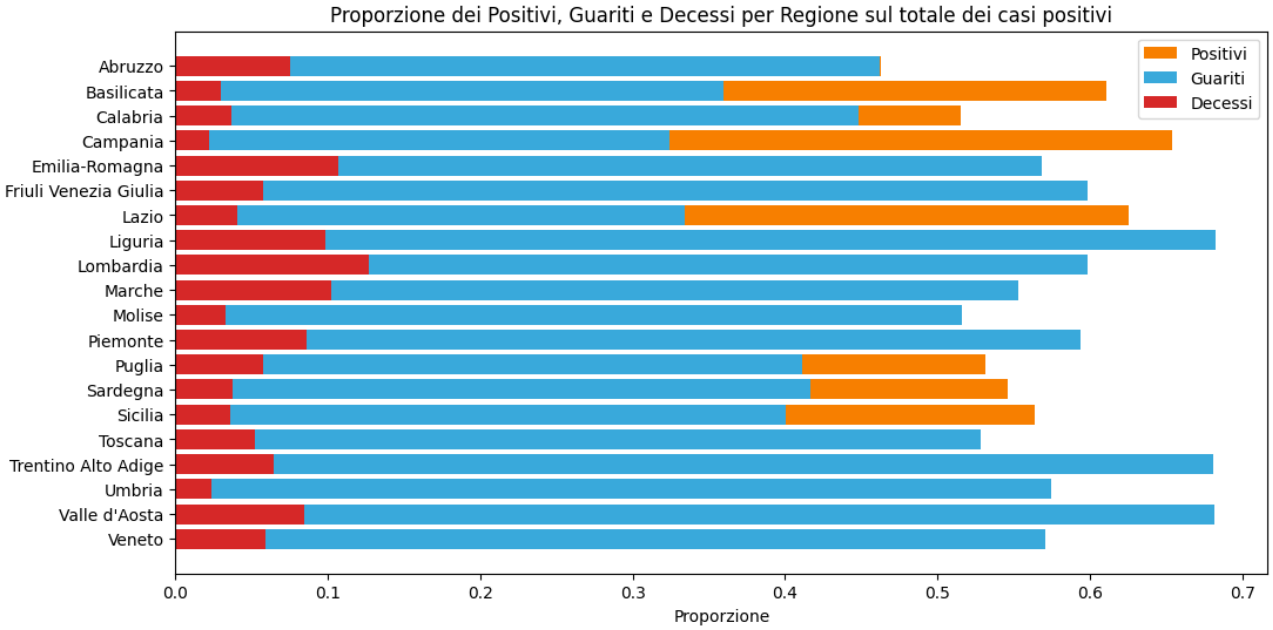
Per poter fare un'analisi più esaustiva, volta a comprendere meglio questi dati, si dovrebbero rapportare le ospedalizzazioni al numero di posti letto realmente disponibili nei reparti intensivi di ogni regione.



Si è poi deciso di analizzare i dati della colonna dei "TotalPositiveCases".

Sulla base della struttura del dataset, essendo il risultato di valori aggregati, si è voluto approfondire questo aspetto e calcolare quanto le singole variabili incidessero sul totale.

Lo studio dell'incidenza del numero di guariti, dei positivi correnti e dei decessi sul totale mostra come i primi costituiscano la maggior parte dei dati presenti nel totale. Questa analisi dimostra anche che, per la struttura del dataset analizzato, per analizzare l'andamento dell'epidemia, sia più corretto utilizzare i dati sui nuovi casi e non il totale dei casi positivi.



3. Analisi per province

Si è analizzato anche il dataset delle province con l'obiettivo di condurre un'analisi più capillare sul territorio italiano. Nello specifico, ci si è chiesto se le province più colpite fossero anche nella regione più colpita dal Coronavirus, ovvero la Lombardia. Le province più colpite sono, in maggior parte, situate in Lombardia o comunque nel Nord-Ovest Italia, in linea con i dati trovati finora.

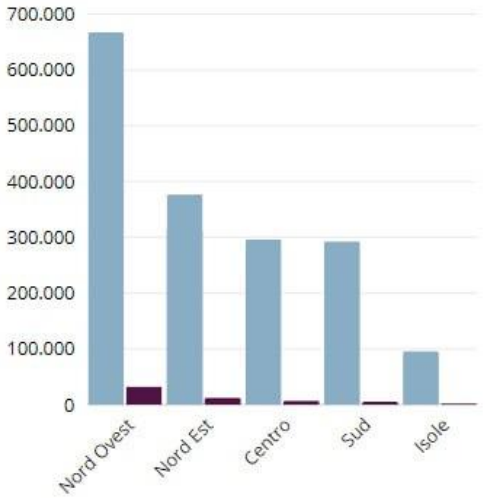
Dall'analisi delle ripartizioni geografiche, si è visto come la zona più colpita (sia per decessi, in viola, sia per contagi, in celeste) quella del nordovest, mentre le isole sono le meno colpite.

4. Comuni

L'obiettivo delle analisi di seguito è stato quello di studiare il tasso di incidenza dei contagi nei singoli comuni. La popolazione è stata suddivisa in diversi scaglioni, rappresentativi di tre classi di abitanti:

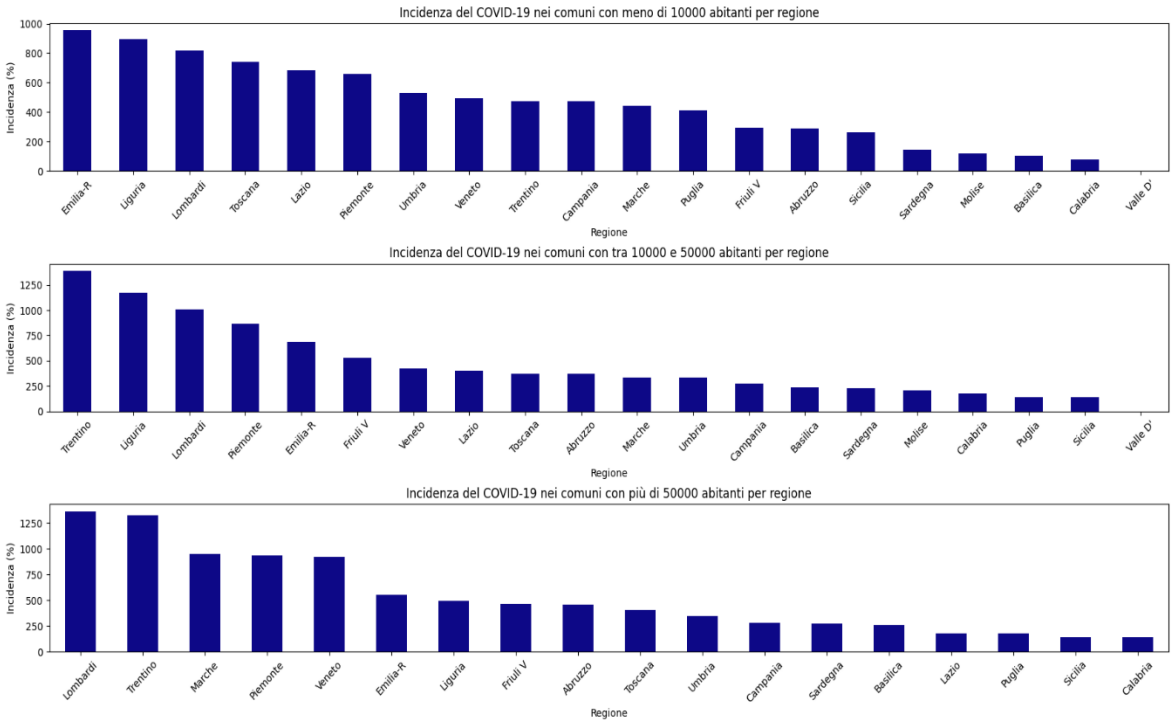
1. Comuni meno popolosi, con meno di 10k abitanti;
2. Comuni mediamente popolati, con popolazione compresa tra 10 e 50k;
3. Comuni altamente popolosi, con più di 50k abitanti.

Province	Tasso di Contagio
Monza e Brianza	5.31%
Varese	5.27%
Aosta	5.23%
Milano	5.22%
Como	5.10%
Bolzano	5.03%



Si nota come in alcune regioni non sono presenti comuni con più di 50k abitanti (Valle d'Aosta e Molise).

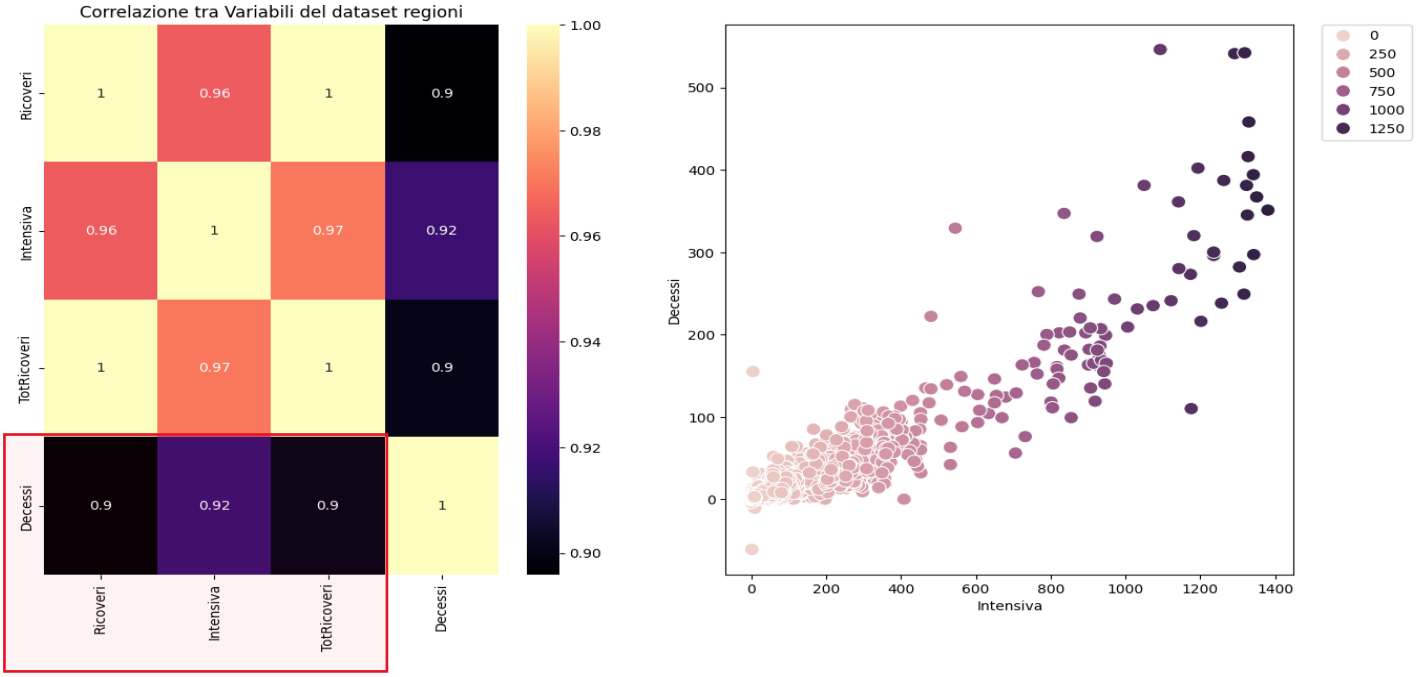
Nei comuni con meno di 10k abitanti, i comuni più colpiti si trovano in Emilia-Romagna, mentre in Lombardia i comuni più colpiti sono quelli con più di 50k di abitanti, seguiti subito dopo dal Trentino.



Correlazioni

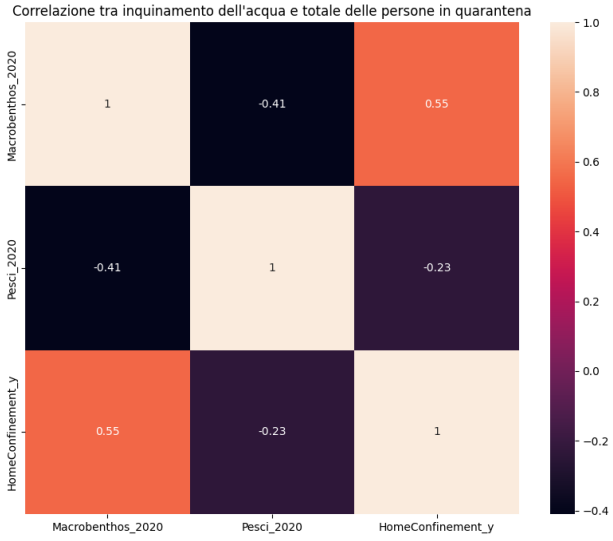
Le analisi di correlazione sono state svolte con l'obiettivo di indagare l'eventuale presenza di relazioni tra i dati presenti nel dataset, in particolare una relazione direttamente proporzionale tra decessi e pazienti ospedalizzati, per cui all'aumentare dei ricoveri ci si aspetta anche un aumento dei decessi.

Dal grafico di sotto si può notare come ci sia una forte correlazione tra i decessi e i ricoveri in ospedale, una relazione ancora più forte per i ricoveri nei reparti di terapia intensiva.

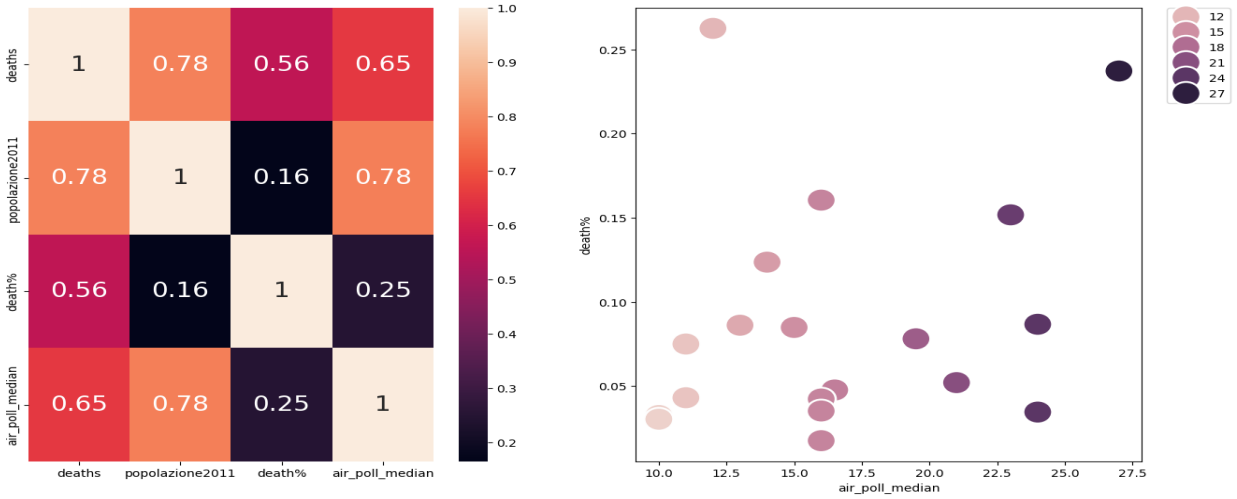


Sono state condotte analisi di correlazione tra la pandemia del Coronavirus e l'inquinamento dell'acqua e dell'aria.

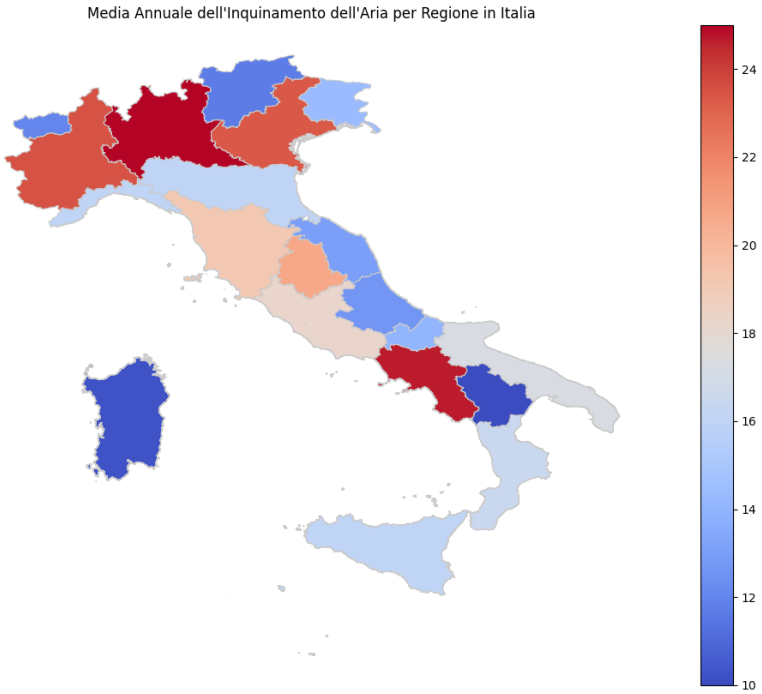
Come dataset sull'inquinamento dell'acqua è stato preso (…). I valori rilevati fanno riferimento ad acqua dolce (fiume). Purtroppo le indagine non ha rivelato l'estistenza di una correlazione tra l'inquinamento delle falde acquifere e il tasso di contagio del covid.



Si è voluto anche verificare la presenza di eventuali correlazioni tra inquinamento dell'aria e aumento dei contagi. Con un dataset trovato su GitHub[1] riferito ai valori mediani di inquinamento nelle diverse regioni italiane, è stata trovata solo una relaziona moderata tra le città più inquinate e morte per Coronavirus.



Tuttavia, la regione Lombardia, l'unica con un valore superiore a quelli consentiti dalla legge, è la regione che ha subito il maggior numero di decessi attribuiti all'azione del virus

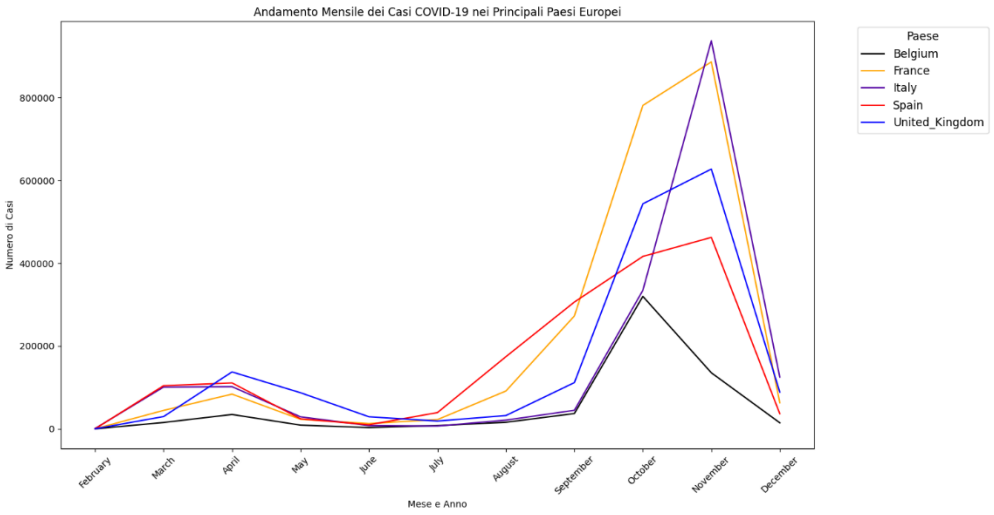


[1] https://github.com/andreapas79/COVID-19/blob/master/df_air.csv

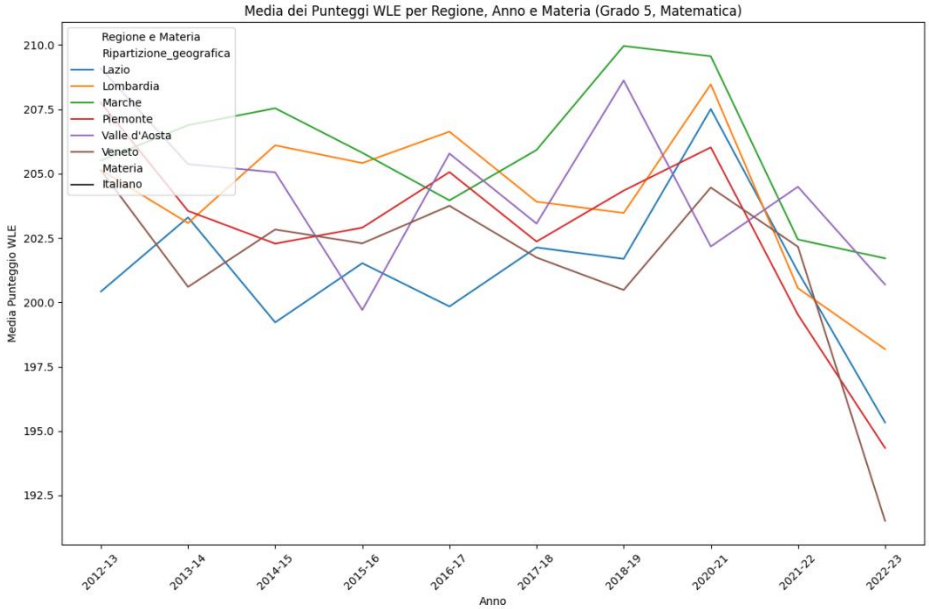
Analisi future

Una volta terminate le analisi sul territorio nazionale, si è voluto investigare l'andamento della pandemia anche negli altri paesi italiani per metterlo in relazione con l'andamento in Italia.

Il dataset è stato reperito sul sito European Data[2]. Da questo sono stati estratti i 25 paesi con un maggior tasso di positività e con un rapporto di popolazione simile a quello italiano. Per questo sono stati selezionati il Belgio, la Francia, la Spagna e il Regno Unito. L'Italia si classifica come prima per numero di contagi, seguita subito dopo dalla Francia, terzo il Regno Unito. Sarebbe interessante approfondire questa analisi e contestualizzarla in base alle misure di gestione dell'epidemia adottate nei singoli paesi.



Un'altra analisi interessante iniziata ma da approfondire riguarda l'impatto della pandemia sull'andamento scolastico, misurato come punteggio alle prove invalsi. Il dataset è stato reperito sul sito ufficiale degli invalsi[3]. Dal grafico sottostante si può notare un calo nei punteggi agli invalsi durante e subito dopo la pandemia. Ulteriori approfondimenti potrebbero riguardare correlazioni tra punteggi e aumento di contagi



[2] https://data.europa.eu/data/datasets/covid-19-coronavirus-data?locale=it
[3] https://www.invalsiopen.it/open-data-invalsi-risultati-2021/

Infine, sarebbe stato interessante analizzare quanto il Covid-19 abbia inciso sulla salute mentale, in diverse fasce d'età e vedere quale sia quella più colpita e se ci siano disturbi mentali associati eventualmente sia come conseguenza diretta del virus sia come effetto secondario.

Gruppo Horror

Federica Branca

Margot Stefanetti

Marco Renato Cerri

Martina Albano